KU LEUVEN

Modeling the dative alternation in four varieties of English

Melanie Röthlisberger & Benedikt Szmrecsanyi

QLVL, KU Leuven

February 24, 2015, Ghent, Belgium

(1) a. He gives $[\text{Mary}]_{\text{recipient}}$ $[\text{roses}]_{\text{theme}}$
(double object construction)

b. He gives $[\text{roses}]_{\text{theme}}$ to $[\text{Mary}]_{\text{recipient}}$
(prepositional dative construction)

→ Semantic equivalence
→ Grammatical acceptable

- Focus on *give*, → e.g. Bernaisch et al. (2014)
- Focus on 1-2 varieties or specific regional varieties, → e.g. Bresnan and Ford (2010), De Cuypere and Verbeke (2013)
- Focus on probabilistic models, → e.g. Bresnan (2007)

→ Combine previous approaches into one
→ Large-scale comparative perspective

1. To what extent do varieties of English share a core probabilistic grammar?
2. Does ecology predict probabilistic similarity between varieties of English – for example, do we find a split between native and non-native varieties of English?

ICE = International Corpus of English

- 1 mio words / 500 texts per variety
- 60% spoken, 40% written
- roughly 12 genres
- 9 varieties: CAN, GB, SIN, IND, NZ, HK, PHI, JAM, IRE
- Focus: CAN, GB, SIN, IND

Extraction

- pos-tagged ICE-corpora
- perl script for extraction of syntactic constructions (verb list of 89 verbs)

Filtering

- 1st filtering - false positives: excluding sentences lacking two overt constituents, non-canonical word order, elliptic structures, coordinated verbs, clausal constituents
- 2nd filtering - choice context: excluding fixed expressions, spatial goals, beneficiaries, ...

Annotation

	Canada	GB	India	Singapore	Total
ditransitive	673 (72.6%)	642 (73.0%)	613 (56.3%)	772 (72.6%)	2,700 (68.2%)
prepositional	254 (27.4%)	237 (27.0%)	476 (43.7%)	291 (27.4%)	1,258 (31.8%)
Total	927	879	1,089	1,063	3,958

Table 1: Distribution of dative variants in four ICE corpora

Restriction to automatic coding

- Variety (GB, CAN, IND, SIN)
- Genre (4 levels: monologues (s), dialogues (s), printed (w), hand-written (w))
- NP expression type (6 levels: common noun ('nc'), proper noun ('np'), personal pronoun ('pp'), impersonal pronoun ('iprn'), demonstrative ('dm'), and gerund ('ng'))
- Length in words of recipient and theme
- Givenness of recipient and theme ('given' vs. 'new')
- Thematicity of recipient and theme
- Type-Token Ratio
- Frequency of recipient and theme

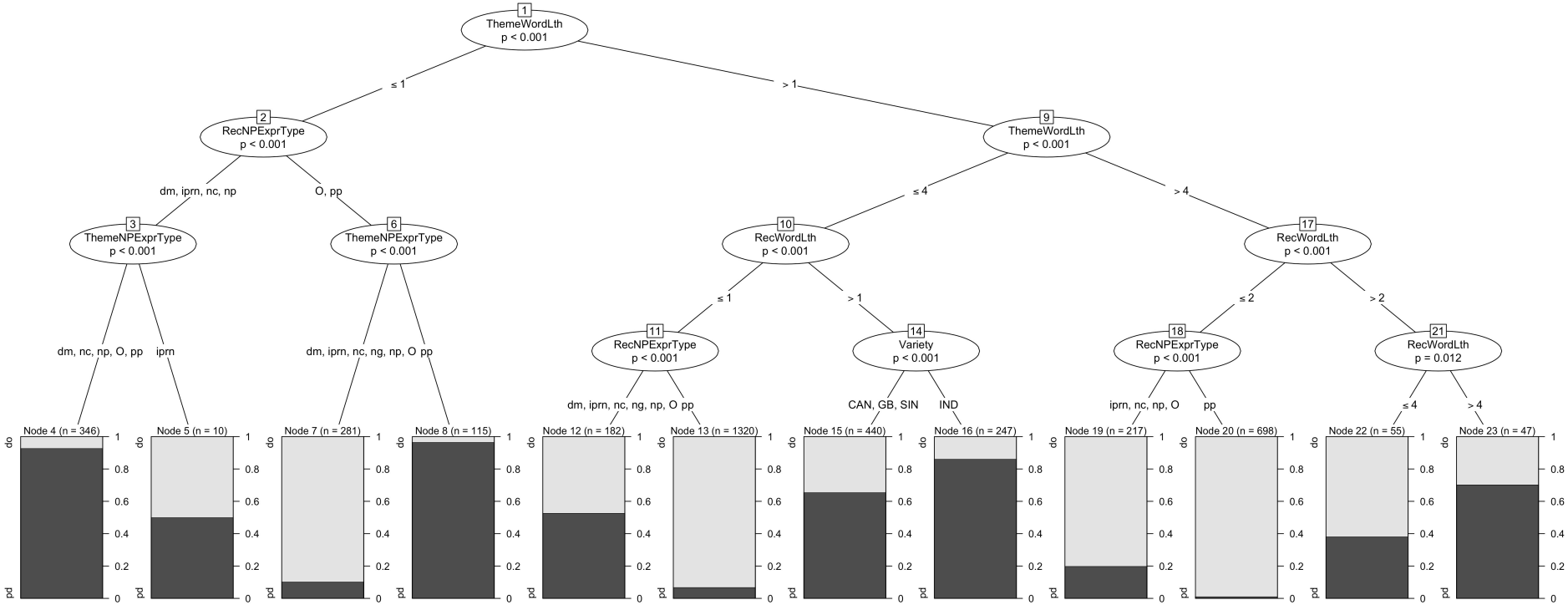
→ additional predictor variables will be added in the next stages

Conditional inference trees : Recursive partitioning/splitting of the data according to the most predictive factor (test of independence); data splits lead to increasingly homogenous subsets wrt levels of response variable → flowchart-like decision tree
disadvantages: new splits are dependent on previous data splits, multicollinearity; see Hothorn et al. (2006)

Conditional random forests : uses randomly sampled subsets of the data to construct trees; permuted variable levels are used to calculate classification accuracy of the tree model (permuted vs. original vs. rest); permuted and original model used for predictions; aggregation of all trees
advantages: good for unbalanced data, empty cells, reduction of collinearity; see Strobl et al. (2008)

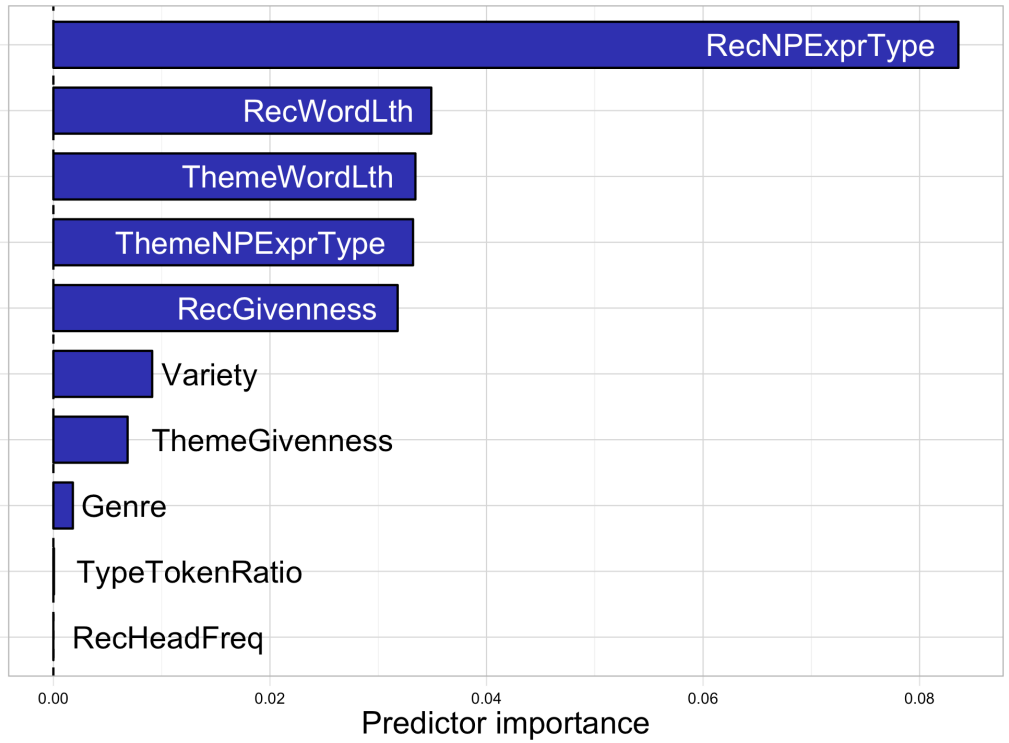
→ R: packages = party (cforest), party kit (ctree)

→ see Tagliamonte and Baayen (2012) for a good explanation of CRF analysis



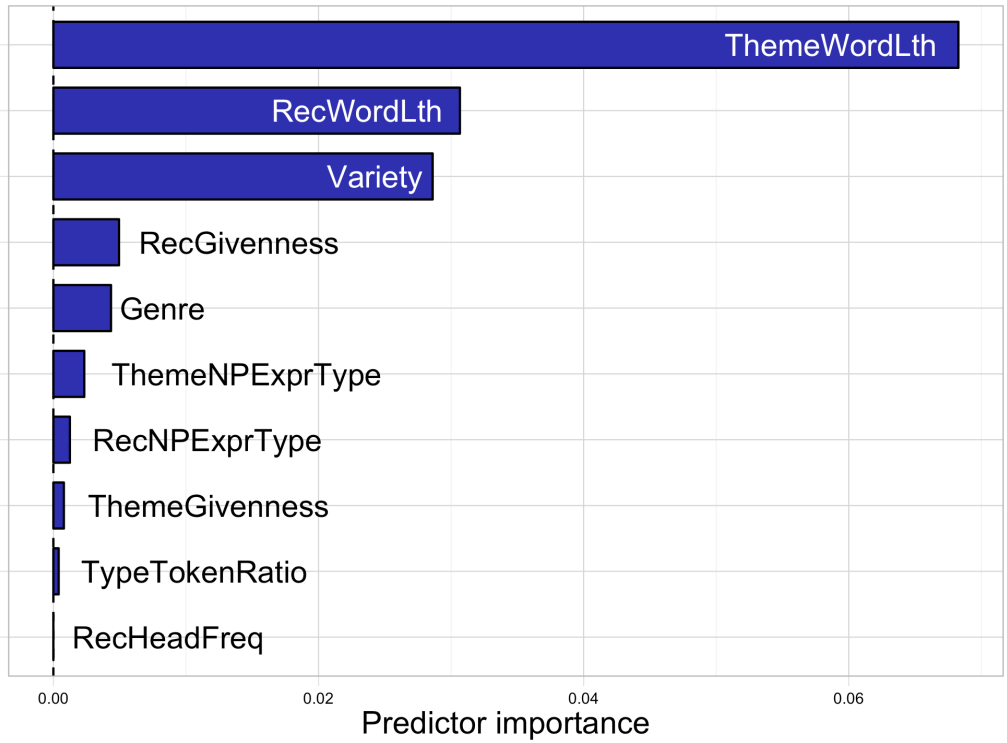
- first tree split according to length of the theme (Node 1)
- if both rec and theme = pronoun → prepositional dative ("give it to him")
- ditransitive favoured when rec = personal pronoun and theme is not (Node 7)
- if both rec and theme are relatively long → prepositional datives (Node 21)
- Variety-based splits: IndE vs. CanE, BrE, SinE

Concordance statistic $C = 0.86$, Classification accuracy = 87.1% (baseline: 68.2%)



- NP expression type of recipient = most important
- Variety / Genre relatively unimportant

Concordance statistic $C = 0.93$, Classification accuracy $= 85.1\%$



Concordance statistic $C = 0.86$

- N=1423
- length of theme = most important
- cross-varietal differences when both recipient and theme are not realised as personal pronouns

1. Conditional inference trees
   - length and expression type = two most important factors → high interaction between pronominality and length
   - principle of end-weight: rec<theme → ditransitive, rec=theme → prepositional
2. Random forests
   - length, pronominality, and givenness of both recipient and theme influence the choice of construction
   - regional variability enters the picture only when pronominal NPs are ignored

1. Do the four varieties share a core probabilistic grammar? Yes.
   - principle of end-weight
   - effect direction of factors are stable across varieties, e.g. pronominal themes favour prepositional dative
   - effect size differs across varieties in those contexts where neither alternate is more or less difficult to process (e.g. non-pronominal)
2. Do we find a split between native and non-native varieties of English? Partially.
   - Indian English is set apart from the other varieties

- hand coding needed for additional predictors, i.e. animacy, semantic class of verb, complexity, ...
- add additional varieties
- mixed-effect logistic regression analysis to account for random effect of "verb"
- cognitive robustness of corpus-derived probabilities will be checked via rating experiments

Thanks!

Bernaisch, T., Gries, S. T., and Mukherjee, J. (2014). The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide*, 35(1):7–31.

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Featherston, S. and Sternefeld, W., editors, *Roots: Linguistics in search of its evidential base*, number 1, pages 75–96. Mouton de Gruyter.

Bresnan, J. and Ford, M. (2010). Predicting Syntax: Processing dative constructions in American and Australian Varieties of English. *Language*, 86(1):168–213.

De Cuypere, L. and Verbeke, S. (2013). Dative alternation in Indian English: A corpus-based analysis. *World Englishes*, 32(2):169–184.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, 9(1):307.

Tagliamonte, S. a. and Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, 24(02):135–178.